The AFRL-MITLL WMT16 News-Translation Task System: We put NMT in your MT Rescoring so you can MT while you MT.

Jeremy Gwinnup[1]**, Timothy Anderson**[20]**,**
Grant Erdmann, Katherine Young[ctr]
Air Force Research Laboratory
first.last.*@us.af.mil

Michaeel Kazi, Elizabeth Salesky,
Brian Thompson
MIT Lincoln Laboratory
first.last@ll.mit.edu

Abstract

This paper describes the AFRL-MITLL statistical machine translation systems and the improvements that were developed during the WMT16 evaluation campaign. As part of these efforts we've adapted a variety new techniques to our previous years' systems including Neural Machine Translation, additional out-of-vocabulary transliteration techniques, and morphology generation.

Preliminary results are denoted with *.

1 Introduction

As part of the 2016 Conference on Machine Translation (WMT16) news-translation shared task, the MITLL and AFRL human language techology teams participated in the Russian–English and English–Russian news translation tasks. Our machine translation (MT) systems represent improvements to both our systems from IWSLT2015 (Kazi et al., 2015) and WMT15 (Gwinnup et al., 2015), the introduction of Neural Machine Translation rescoring, neural-net based recasing, unsupervised transliteration of out-of-vocabulary (OOV) words (Durrani et al., 2014), and an unique selection process for language modelling data (Erdmann et al., 2016). For the English–Russian translation task we experimented with morphology generation techniques to improve translation quality.

2 System Description

We submitted systems for the Russian–English and English–Russian news-domain machine translation shared tasks. In most submission systems, we used either phrase-based or hierarchical variants of the moses decoder (Koehn et al., 2007). In some cases we used the performance-enhanced version of Moses (Hoang et al., 2016). As in previous years, our submitted systems used only the constrained data supplied when training.

2.1 Data Usage

In training our systems we utilized the following corpora to train translation and language models: Yandex[1], Common Crawl (Smith et al., 2013), LDC Gigaword English v5 (Parker et al., 2011) and News Commentary. For additional language modelling data we processed the new Common Crawl monolingual corpora using the techniques described in Erdmann et al. (2016).

The Wikipedia Headlines corpus[2] was reserved to train named entity recognizers.

2.2 Data Preprocessing

We processed the training data similarly to our WMT15 system (Gwinnup et al., 2015). We also examined irregular behaviors in Moses's punctuation normalization script[3]. We also run a script that examines the source and target side of the parallel training data and removes lines that are identical in both the source and target in order to prevent the effects of wrong-language phrases "polluting" the phrase and rule tables.

2.3 Phrase Table Generation

We used the standard Moses method of extracting and creating phrase tables. Phrase tables were binarized using either the Compact Phrase Table (Junczys-Dowmunt, 2012) or ProbingPT (Hoang et al., 2016) methods.

This work is sponsored by the Air Force Research Laboratory under Air Force contracts FA-8721-05-C-0002 and FA-8650-09-D-6939-029.

[1] https://translate.yandex.ru/corpus?lang=en

[2] http://statmt.org/wmt15/wiki-titles.tgz

[3] normalize-punctuation.perl

2.4 Tuning Improvements

Improvements were made to our tuner, Drem (Erdmann and Gwinnup, 2015), since our last submission. Rescoring weights are now not penalized in the n-best list interpolation scheme, since they do not directly affect n-best lists. This new feature provides faster convergence of our NMT-rescored systems. Another improvement to Drem is that the metric chrF3 (Popović, 2015) is now available as a tuning objective function. It was applied in the English–Russian experiments.

2.5 Neural Network Recaser

We noticed a substantial gap between uncased and cased BLEU scores on our systems. Attacking the problem in post-processing, it became apparent that recasing can only do so much on monolingual data. We therefore built a classifier that uses both the source-side and the target-side of the translations. The inputs to the classifier are:

- t_i, the word to be recased, as well as t_{i-1} and t_{i-2}
- $s_{a(i)}$, the source word aligned to t_i, plus $s_{a(i)\pm 1}$. Alignments were taken from Moses output, and missing alignments were computed using the NNJM affiliation heuristic (Devlin et al., 2014).
- The status of the source word as lowercase, capitalized, or OTHER.

The exact classifier used could be anything, but a neural network is simple to create and robust. Our architecture is as follows:

1. Vocabulary of all words, excluding 25% of singletons
2. Input: Word vectors for these words, plus nine binary inputs ($s_{i-1} = lc, s_{i-1} = Uc, s_{i-1} = OTHER, s_i = lc \ldots$), all concatenated together into a single vector
3. Two hidden layers, default size 100
4. One softmax output, 3 output classes

The resulting recaser consistently yields +0.2-0.25 case-sensitive BLEU over a standard language model recaser.

2.6 Inflection Generation

English-Russian systems have the added challenge of generating morphologically rich word-forms. In addition to an English-Russian baseline, we trained two method to generate inflected forms. First, we trained an MT system from English to lemmatized Russian, using the Mystem parser to lemmatize all available parallel data, and then trained a MT system from lemmatized Russian to Russian. Scoring against lemmatized references, the first step yielded 27.70 case-insensitive BLEU on `newstest2016`. However, while the lemru-ru system was successful with one-to-one lemmatized training data, it couldn't recover from mistakes in the MT output of the first step and the system overall did not perform as well as our baseline (17.19* cased BLEU).

Original: Woud n't you know it ?
Annotated: Would n't you know-2p it ?
Dependency Parse:

Index	Word	POS	Head	Relation
1	Would	MD	4	aux
2	n't	RB	4	neg
3	you	PRP	4	nsubj
4	know	VB	0	root
5	it	PRP	4	dobj
6	?	.	4	punct

Figure 1: Annotation via Dependency Parse

We also attempted to address inflection generation during training using verb annotation, following the approach of Kirchhoff et al. (2015) for Arabic verb inflection. We use dependency parsing to identify the subject of the verb in the English sentence and then annotate the verb with the person, number, and gender characteristics of the subject. This provides the potential for the system to match annotated English verbs to the correctly inflected Russian verbs during training. Figure 1 shows an annotated sentence and the underlying dependency parse.

We use the Stanford parser (Klein and Manning, 2003) and conversion utility to generate the dependency parses, adjusting the tokenization of the input to match the Stanford treatment of contractions. We apply annotation to verbs with subjects listed as *nsubj* or *xsubj* in the dependency parse. Person, number, and gender are derived from the subject's POS tag and from the specific lexical item for pronouns. Coordinate subjects are counted as plural.

An unannotated MT system has a good chance of associating the correct verb form with the sub-

Would n't **you know**-2p it ?
The **country** was gradually **recovering**-3p-sg ..
The **interests** of people **take**-3p-pl precedence ..

Figure 2: Annotation at different separation dists.

ject if the subject and verb are adjacent and can be extracted as a phrase, while more distant pairs are less likely to be found in the phrase table, leaving the verb open to translation in the wrong inflected form. Since annotation can increase data sparsity, it is better to apply it only when necessary.

Kirchhoff et al. (2015) address the data sparsity issue by only applying their annotation-trained model when their baseline model translates the subject and verb via separate phrases. In our system, we simulate the use of a backoff model by restricting our annotation to subjects and verbs that occur with a minimum separation distance.

Figure 2 shows the effect of specifying a minimum separation distance. In the first sentence, the subject and verb are adjacent; any separation requirement greater than zero prevents annotation of the verb. The other sentences show a greater separation, and annotation will be maintained if the separation requirement is less than 3.

We also created a factored version of the verb annotation system to avoid the data sparsity problem. The annotations were specified as factors on the verb, with a null factor on the unannotated words, e.g. `would|NONE n't|NONE you|NONE know|2p it|NONE ?|NONE`

2.6.1 Discussion

We use a corrected version of the Hjerson (Popović, 2011) error analysis program to examine the effect of inflection generation on the Russian output. We found that technique alpha provided the most improvement for inflected forms. Table 1 shows inflection errors as percent of hypothesis words for each method.

Technique	Inflection Errors	BLEU
baseline	20%*	23.10*
alpha	14%*	23.56*
beta	17%*	23.35*

Table 1: Hjerson performance

We revised Hjerson to report the specific inflection errors, and used Mystem to analyze the affected POS and morphological information. We found that technique alpha improved the inflection of verbs more than nouns.

2.6.2 Transliteration

We employed multiple methods to address transliteration of remaining out-of-vocabulary (OOV) words: our method of selectively transliterating OOVs as outlined in last year's work, an unsupervised statistical transliteration approach and a novel character-based neural-network tranliteration approach. We also experiment with combinations of these three techniques.

Selective Transliteration

We created a list of 54k Named Entity (NE) pairs from the Common Crawl using transliteration mining (Gwinnup et al., 2015) and used this list to translate OOVs. Remaining OOVs were selectively transliterated, using capitalization of the Russian source as an indication of a NE that should be transliterated; remaining OOVs were dropped.

Unsupervised Statistical Transliteration

As a contrast to our selective transliteration approach, we also experimented with using the unsupervised statistical transliteration method (Durrani et al., 2014) included in Moses.

Neural Network Transliteration

We built a neural network based transliterator using the 54k mined transliteration pairs described above. We trained an encoder-decoder LSTM network to produce characters in a target language given characters from a word in the source language. The network configuration was nearly the same as that in our NMT experiments, except the network was significantly smaller (hidden sizes of 100 and 200, with 1, 2, and 3 hidden layers) and had a beam of 5. A small (5k) subset of the data was held out for evaluation/tuning. Since Russian nouns use case, multiple Russian word forms may map to a single English spelling. For this reason, we tried rescoring with a unigram language model trained on the monolingual data to help weight the correct English spelling of words that may have been seen in the language modeling data but were not in the phrase table. The LM's unknown word probability was optimized on the validation set.

We integrated this into our SMT pipeline through different backoff phrase tables. Unknown words from the dev and test sets were mapped to their top 10* best transliterations from the final

System	Exact matches
Baseline [0 edit distance]	23.1%
Single enc-dec	34.7%
Ensemble (6)	38.7%
Single enc-dec + LM rescore	42.5%
Ensemble (6) + LM rescore	45.8%

Table 2: Fraction of transliterations that match exactly, on validation set (subset of newstest2014)

system in Table 2 to create phrase table entries. The results are in Table 3.

System	Cased BLEU
1. drop unknowns	28.30
2. pass-through unknowns	28.10*
3. ascii entries in backoff PT	28.15*
4. 3 + cased words transliterated	28.47*
5. 3 + all Cyrillic transliterated	28.45*

Table 3: Neural Transliteration via Backoff PTs

2.7 Neural MT

We describe a Neural Machine Translation system we developed and our strategies to integrate this system into our machine translation framework.

2.7.1 System

We trained a neural encoder-decoder network (Sutskever et al., 2014; Bahdanau et al., 2014; Luong et al., 2015) using the attention model from (Vinyals et al., 2015) to perform neural machine translation (NMT). We trained the model using Adagrad (Duchi et al., 2011) and found it improved performance over the learning rate schedule proposed in (Luong et al., 2015). We also found it advantageous to use a larger source vocabulary (200k-500k words worked well). Each had two 1000-dim hidden layers, with beam and stack of 5. Our NMT results are shown in Table 4. They did not perform competitively with our SMT systems by themselves, however they were very useful in rescoring, and system combination.

System	Cased BLEU
1. Single model	21.00*
2. Ensemble of 2	21.46

Table 4: Russian–English Neural MT Systems decoding `newstest2015`

2.7.2 Reranking

We compared two different ways of using the NMT system to augment our phrase-based system.

1. **Single set of weights** We augment the Moses n-best list with NMT scores for each sentence, and then tune the decode weights using Drem. We repeat this process 10 times, using the last weights to decode the test setand one-best calculation.
2. **Decode + rerank weights** We tune the decode weights using Drem, without the NMT scores. After 10 iterations, we merge the n-best lists together and compute NMT scores over the result. Then, we compute a second set of weights. To decode the test set, we pass the decode weights to moses, augment the n-best list with NMT scores, and finally apply the one-best dot product using the second set of weights.

The first process produced scores of 27.50, and the second 28.10 (mteval, case+punc, tst2015, average of 6).

2.8 NNLM Decoding

We also experimented with neural network language model decoding. Seeing the gains produced last year at WMT (Alkhouli et al., 2015), and the decent perplexities achieved by RNNs, we integrated RNNLM into Moses and used the self-normalization speed-up proposed by (Devlin et al., 2014), with α, the term that measures the deviation on the output layer from summing to 1, set to 0.1. In all experiments, we allowed the decoder to recombine hypotheses that matched the previous 8 target words. Moses was run with 24 threads. We summarize the results in Table 5. Networks were trained with Theano, on a sentence-by-sentence basis, both to match the decode-time use case, and to avoid backpropagation through time.

3 Results

We submitted 3 Russian–English and 4 English–Russian systems for evaluation, each employing a different decoding strategy. Each system is described below. Automatically scored results reported in BLEU (Papineni et al., 2002) for our submission systems can be found in Table 6 for Russian–English and Table 7 for English–Russian.

Finally, as part of WMT16, the results of our submission systems listed in Tables 6 and 7 were

Model	Perplexity	Decode time (s)	BLEU
0. Baseline	–	300*	27.30*
1. RNN 500-dim	160*	1700	27.30*
2. RNN 800-dim	140*	2700*	27.40*
3. LSTM 500-dim	120*	3500*	27.45*
4. LSTM 2x500-dim	95*	8000*	27.50*

Table 5: Russian–English MT decoding newstest2015 with recurrent neural language models

System	Cased BLEU	Uncased BLEU
1. pb-lc+hierLR+TCNN+NMT-rescore	28.18*	29.21*
2. NMT 2x1k ensemble2 Tensorflow	21.00*	–
3. (1) + (2) combination (Rosti et al., 2008)	28.31*	29.40*

Table 6: Russian–English MT Submission Systems decoding newstest2015

System	Cased BLEU	Unc. BLEU
enru-pb	23.42	23.52

Table 7: English–Russian MT Submission System decoding newstest2016

ranked by monolingual human judges against the machine translation output of other WMT16 participants. These judgments are reported in WMT (2016).

4 Conclusion

In conclusion, we present a series of improvements to our Russian–English and English–Russian machine translation systems which represent significant increases in machine translation quality.

References

Tamer Alkhouli, Felix Rietig, and Hermann Ney. 2015. Investigations on phrase-based decoding with recurrent neural network language and translation models. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 294–303, Lisbon, Portugal, September. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Proceedings of the ACL, Long Papers, pages 1370–1380, Baltimore, MD, USA.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Nadir Durrani, Hassan Sajjad, Hieu Hoang, and Philipp Koehn. 2014. Integrating an unsupervised transliteration model into statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 148–153, Gothenburg, Sweden, April. Association for Computational Linguistics.

Grant Erdmann and Jeremy Gwinnup. 2015. Drem: The AFRL submission to the WMT15 tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 422–427, Lisbon, Portugal, September. Association for Computational Linguistics.

Grant Erdmann, Jeremy Gwinnup, and Timothy Anderson. 2016. Sampling “good” data from a “ridiculous” sized corpus for language modeling. In *Proceedings of the First Conference on Statistical Machine Translation*, Berlin, Germany, August. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning.

Jeremy Gwinnup, Tim Anderson, Grant Erdmann, Katherine Young, Christina May, Michaeel Kazi, Elizabeth Salesky, and Brian Thompson. 2015. The AFRL-MITLL WMT15 system: There's more than one way to decode it! In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 112–119, Lisbon, Portugal, September. Association for Computational Linguistics.

Hieu Hoang, Nicolay Bogoychev, Lane Schwartz, Kenneth Heafield, and Marcin Junczys-Dowmunt. 2016. Fast, scalable phrase-based SMT decoding. In *Proceedings of the 54th Annual Conference of the Association of Computational Linguistics (ACL 2016)*, Berlin, Germany, August. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2012. Phrasal rank-encoding: Exploiting phrase redundancy and translational relations for phrase table compression. *Prague Bulletin of Mathematical Linguistics*, 98:63–74.

Michaeel Kazi, Brian Thompson, Elizabeth Salesky, Tim Anderson, Grant Erdmann, Eric Hansen, Brian Ore, Katherine Young, Jeremy Gwinnup, Michael Hutt, and Christina May. 2015. The MIT-LL/AFRL IWSLT-2015 systems. In *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT'15)*, Da Nang, Vietnam, December.

Katrin Kirchhoff, Yik-Cheung Tam, Colleen Richey, and Wen Wang. 2015. Morphological modeling for machine translation of english-iraqi arabic spoken dialogs. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 995–1000, Denver, Colorado, May–June. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pages 311–318, Philadelphia, Pennsylvania, July.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition LDC2011T07. *Philadelphia: Linguistic Data Consortium*.

Maja Popović. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 96:59–68, 10.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Antti-Veikko I Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2008. Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 183–186. Association for Computational Linguistics.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL '13)*, pages 1374–1383, Sofia, Bulgaria, August.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pages 2755–2763.

WMT. 2016. Findings of the 2016 Conference on Statistical Machine Translation. In *Proceedings of the First Conference on Statistical Machine Translation (WMT '16)*, Berlin, Germany, August.